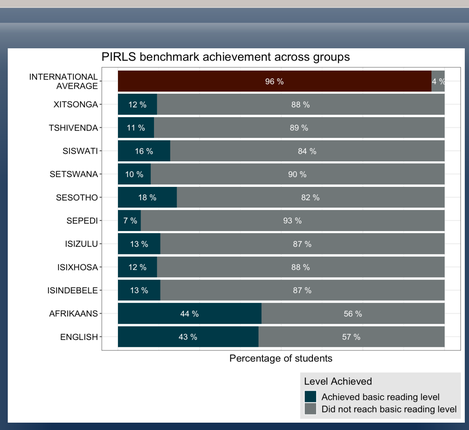
The overall impact of cross-language Differential Item Functioning at the test level: PIRLS 2016 in South Africa

Heather Leigh Kayton | PhD candidate
University of Oxford, Department of Education

PROBLEM

In South Africa's linguistically diverse landscape, students take PIRLS (an international large-scale reading assessment) in one of 11 official languages. However, South African students score well below the international average, and substantial reading achievement gaps exist between language groups within the country. Added to this, PIRLS results feature strongly in policy and intervention conversations in South Africa. It is therefore essential to ensure PIRLS measures reading abilities fairly across languages.



Only 7% of Sepedi students reached the basic reading level on PIRLS compared to 43% of English students, and 96% internationally

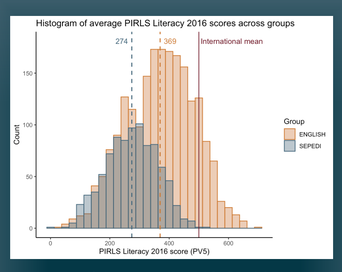
DATA

Two language groups (N=2,987) were selected for comparison from the South African national sample:

- In Grade 4 (~10 years)
- Attended English or Sepedi medium schools
- Took PIRLS Literacy 2016 in English (n=2,089) or Sepedi (n=898)

PIRLS Literacy 2016:

- Reading comprehension assessment
- 183 items across 12 reading passages
- Due to targeting concerns, only items that met basic model and targeting requirements were analysed (152 of 183)
- 79 multiple choice, 73 constructed response



The English group (M 369, SD 115) had significantly ($p < .01$) higher average score than the Sepedi group (M 274, SD 85.9)

DIF happens, but how does it impact comparability?

METHODS

RQ: To what extent does the item-level comparability of the English and Sepedi versions of PIRLS 2016 impact on the comparability of the assessment at test level?

ITEM LEVEL IMPACT

- IRT-based differential item functioning (DIF) analysis
- IRT model (2PL & GPCM)
- Likelihood-ratio tests to test for DIF, IRT-LR (Thissen, 2001)
- Effect sizes determined using Meade's (2010) taxonomy

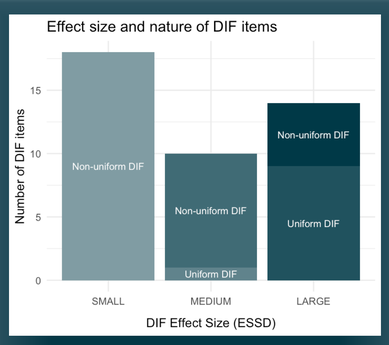
TEST LEVEL IMPACT

- Differential response functioning (DRF) framework (Chalmers, 2018)
- DRF quantifies cumulative effect of DIF to determine test level impact
- Used signed (directional) and unsigned (absolute) DRF statistics

DIF AT ITEM LEVEL

DIF occurs when students from different groups with the same overall ability have a different probability of answering an item correctly

- 42 items (28%) showed significant DIF
  - 14 large
  - 10 medium
  - 18 small
- Majority of small/medium DIF items were non-uniform
  - This means they advantage different groups at different levels of ability
- Higher number of large uniform DIF items advantaged the Sepedi group (5)



DIF findings point to issues with item comparability across languages, but more is needed to understand what this means at the overall test level

TEST LEVEL IMPACT

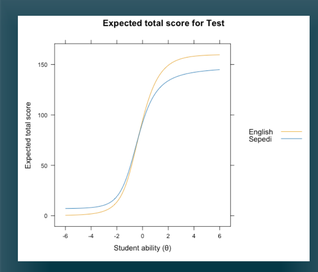
Statistic	Value	2.5% CI	97.5% CI	χ^2	df	p-value
sDRF	-1.24	-3.7	.81	1.19	1	.275
uDRF	5.29	4.15	7.26	72.97	2	< .001*

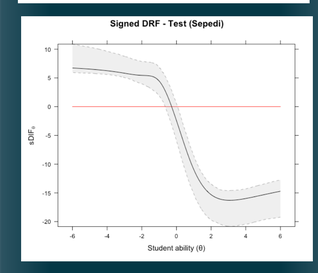
Unsigned DRF (uDRF = 5.29, 97.5% CI: 4.15 to 7.26)

- Significant absolute divergence between expected scores across groups ($p < .01$)
- This substantial uDRF indicates an average absolute difference of 5.29 points between the English and Sepedi versions.

Signed DRF (sDRF = -1.24, 97.5% CI: -.81 to 1.19)

- No significant directional divergence in expected scores across groups
- This non-significant result indicates the presence of multidirectional (non-uniform) DIF effects that cancelled each other out at the overall test level





While no consistent directional bias was found, significant comparability issues exist overall between the English and Sepedi versions

IMPLICATIONS

- Direct score comparisons between English and Sepedi groups are problematic given evidence of expected score divergence for students of equivalent ability across groups.
- Actual gaps may be obscured by multidirectional DIF; true differences are potentially larger.
- Validity concerns arise regarding the construct being measured in each of these languages.
- Rigorous within-country comparability evaluations are vital for fair ILSA interpretations.
- Findings underscore risks of using PIRLS cross-language results to guide interventions and policy.

Ensuring comparability across all language groups in future adaptations should be a priority.

Heather Leigh Kayton
heather.kayton@education.ox.ac.uk

DEPARTMENT OF EDUCATION
UNIVERSITY OF OXFORD